# Prediction of Stroke using Data Mining Classification Techniques

Ohoud Almadani, Master of Health Informatics (MHI), and Registered Pharmacist (R.Ph)
Pharmacutical care department at King Abdulaziz Medical City
Riyadh, KSA

Riyad Alshammari
King Saud bin Abdulaziz University for Health Sciences
King Abdullah International Medical Research Center (KAIMRC)
Minstry of National Gurad Health Affairs
Riyadh, KSA

***Abstract*—Stroke is a neurological disease that occurs when a brain cells die as a result of oxygen and nutrient deficiency. Stroke detection within the first few hours improves the chances to prevent complications and improve health care and management of patients. In addition, significant effect of medications that were used as treatment for stroke would appear only if they were given within the first three hours since the beginning of stroke. A framework has been designed based on data mining techniques on Stroke data set that is obtained from Ministry of National Guards Health Affairs hospitals, Kingdom of Saudi Arabia. A data mining model was built with 95% accuracy. Furthermore, this study showed that patient with the following medical conditions, such as heart diseases (hypertension mainly), immunity diseases, diabetes militias, kidney diseases, hyperlipidemia, epilepsy, or blood (platelets) disorders has a higher probability to develop stroke.**

***Keywords*—*Stroke; data mining; classification***

## I. Introduction

Knowledge Discovery from Data (KDD) is a growing field of computer science that deals with information gain and decision support through large data analysis and automated extraction of patterns.

Information gain from health data may lead to innovative solution or better treatment plan for patients. In order to gain knowledge intelligently from stroke data, a data mining technique is utilized to semi-automatically process data and generate data mining model that can be used by health care professionals [1].

A stroke is a neurological disease that occurs when a brain cells die because of oxygen and nutrient deficiency. Occlusion of brain blood vessel by a clot or blood vessel rupturing are the major causes of oxygen and nutrient supply deficiency [2]. Cerebro-Vascular Accident (CVA) is the previous name of stroke, which divided nowadays into three types known as Hemorrhagic stroke, Acute Ischemic stroke, or Transient Ischemic Attack [2], [3].

Stroke detection within the first few hours improves the chances to prevent complications and improve health care and management of patients [4]. In addition, significant effect of medications that used as treatment for stroke will appear only if they were given within the first three hours since the beginning of stroke [4].

According to heart disease and stroke statistics update of 2015 [5], 11.13% of deaths globally were accounting for stroke. With 33 million affected persons, stroke is the second leading cause of death worldwide. Furthermore, it is the first leading cause of adult disability with 16.9 million affected persons [5], [6].

According to the Global Burden of Diseases (GBD), Disability-Adjusted Life Years (DALYs) measure showed that the rate of DALYs in Saudi Arabia increased by 50% from 1990 to 2010 because of stroke [6], [7]. In addition, the number of Years of Life Lost (YLLs) statistics of 2010 showed that stroke was the fourth reason of death in Saudi Arabia with an increased rate of 52% since 1990 [8]. Therefore, early prediction of stroke will facilitate effective therapy administration within an appropriate period [9].

The main objectives of this research are twofold: i) Use data mining techniques to predict patient at risk of developing stroke; and ii) Find the patient with who has higher chances to develop stroke. Therefore, three classification algorithms, namely C4.5, Jrip, and multi layers perceptron (MLP), are used on stroke patient data set collected from National Guard hospitals in three different cities in Kingdom of Saudi Arabia. The three classifiers are compared with each to find the best the performance with the goal of finding the best predication model. Hence, framework has been developed to identify stroke patients using proper decision support tool that would help in achieving the following goals: i) Decrease the impact of stroke on patient life; ii) Improve country's population life expectancy and health; and iii) Reduce health care budget.

The remaining segments of this research article are arranged as following: Section 2 presents the literature review on using data mining to predict stroke. Section 3 explains the methodology while Section 4 presents the results and discussion. Finally, Section 5 includes the conclusion and future work.

## II. Literature Review

Stroke has a high impact on public health and countries' economies that lead to build several stroke associations with the aim of improving lives quality by providing public health education, lifestyle modification, evidence-based treatment guidelines, and CardioPulmonary resuscitation (CPR) training [5]-[9]. In addition, it leads to conduct multiple researchers, which focus on finding preventive and educational materials for stroke [5]-[9]. Defining risk factors were the main goal for several researches about stroke [10]-[21]. A research was conducted in Taiwan, that showed age had a significant risk factor for stoke with patients older than 65-year-old with hypertension, and diabetes mellitus (DM), while, gender and cerebral ischemic events were non-significant factors [10]. Another study took its place in United States revealed that the main risk factors were hypertension, DM, hyperlipidemia, smoking, obesity, and congestive heart failure [11].

With the advance development of technology and the high performance of data analysis tools, health care researchers seek a suitable tool to prevent or detect acute stroke in its early stages. Data mining technique provides researchers with a helpful tool to analyze a large amount of data, such as in the case of health care organization, and facilitate the detection of common patterns for such conditions. Therefore, it could provide a prediction model to identify possible individuals to develop such disease [12]-[18].

Decision support tools were the main outcome for many health-related data mining articles. Sheng-Feng Sunga, et al. [19] analyzed data of acute ischemic stroke patients to develop a prediction model for the severity of the disease. In their study, they used K-nearest neighbor model, multiple linear regression, and regression tree model, that resulted an accuracy of 0.743, 0.742, and 0.737, with 95% confidential interval [19].

Ahmet K. Arslan et al. [20] used three data mining algorithms, namely: Support Vector Machine (SVM), Stochastic Gradient Boosting (SGB) and penalized logistic regression (PLR) to predict stroke. SVM achieved an accuracy of 98% [20]. In addition, by using K-nearest neighbor and C4.5 decision tree, Leila Amini et al. [18] achieved an accuracy of stroke prediction equal to 94.2% and 95.4% respectively. Artificial Neural Network (ANN) prediction model achieved a predictive accuracy of thrombotic stroke equal to 89% as shown in Shanthi et al. study [21]. Stroke is being observed as a rapidly growing health issue in Saudi Arabia. It is the second cause of death by killing 14.4 thousand people in 2012. Therefore, it becomes one of the health care issues in Saudi Araba. The lack of researches that focus on the role of technology, mainly KDD, in predicting of stroke in the Saudi Arabia, leads to this research.

## III. Methodology

In this section, the methodology is explained including on how the data sets are obtained, attributes, the data mining algorithms and the evaluation criteria.

### A. Data Collection

Data received from the data governance department at King Abdulaziz Medical City (KAMC). KAMC opens on 2001 at Riyadh city. KAMC grows up to become one of the top hospitals in the Middle East with a bed capacity of more than 1500 by 2016. KAMC is serving 2.5 million outpatients and around 60,000 in-patients annually. The data set was extracted from KAMC contained all patients who were diagnosed as stroke case or stroke mimic case on 2016 from the 2nd of January to 31st of September. The reason behind using this time frame is due to the installation of new Health Information System at KAMC. There are two classes in the data set. The first class includes the medical records for patient's known to have stroke while the second class includes records of stroke mimic patients, who usually misdiagnosed as stroke patient due to the similarity of the symptoms. This data set consists of 969 instances, 69 of them classified as stroke mimics while 899 classified as stroke patients. The data set contains 360 females (37.15 %) 33 of them diagnosed as stroke mimic and 327 as stroke cases. As well, 607 males (62.6%) 36 of them are stroke mimic while 571 are stroke cases (Fig. 1).

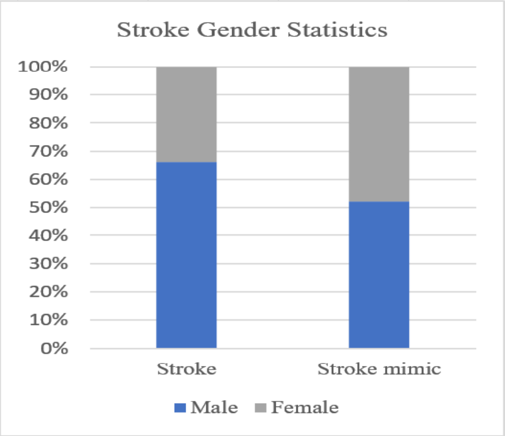

Fig. 1. Stroke gender statistics.

### B. The Attributes

The obtained data set contained 1004 attributes. It main attributes are the class (stroke, stroke-mimic), age (ordinal), gender (female, male), number of medication (numeric), medication name (taken, not-taken), and lab test name (normal, abnormal).

Attribute selection was applied then to reduce data dimensionality. Attribute selection is a data mining technique that used to select the most relevant attributes [1], [22]. Principle Component Analysis (PCA) is utilized [22]. It is a data mining technique that works on dimensionality reduction. The idea of this technique is to create a new alternative attribute that combines several previous attributes essence. This algorithm has the ability of reducing attribute noise by transforming it to the PCA space, eliminating the worst eigenvector then transform it back to the original space [1], [22].

The final attribute sets are related to patient age, gender, lipid disorder, lab test abnormalities, hypertension medications, diabetes medications, and other medications are included, which resulted of data set that contains 147 attributes. Moreover, data are divided to two separate data sets: training data set to build the model, and test data set to evaluate the model.

### C. Data Mining Algorithms

For their known high accuracy rate, J48 (C4.5), JRip, and Neural Network (multilayer perceptron [MLP]) algorithms were applied on the stroke training data set to build a model. All Data mining algorithms have been applied using Weka Software (Version 3.8, Machine Learning Group, University of Waikato, Hamilton, New Zealand). C4.5- J48 in WEKA, is an algorithm that works first by choosing the root attribute through attribute selection (gain ratio) [22], [24], [25]. It then works to build decision tree branches from that attribute values and distribute instances into its corresponding branch [22], [24], [25]. This process will be repeated until all instances are assigned to their correct class [22], [24], [25]. On the other hand, RIPPER algorithm, Jrip in WEKA, is a rule based algorithm [22], [26]. The algorithm takes certain steps to build its model. First, a set of rules will be constructed using incremental reduced error [22], [26]. Then each class will be examined against those rules repeatedly until all instances of each class are covered [22], [26]. At the end, rules that cover all classes will be used to build the model [26].

The third algorithm is a Neural Network called Multilayer perceptron (MLP). MLP is a forward feed neural network [22], [27], [28]. It uses one direction feed of input through one or more layers to produce output layer [22], [27], [28]. To train this algorithm a back-propagation learning algorithm usually used, and it helps to solve non-linearity problem [22], [27], [28].

## IV. Results

Results are shown for data with all attributes (row data) and data after attributes section (data after using PCA).

### A. Results of Prediction Model with All Attributes (Raw Data)

The comparison of the data mining algorithms used with 10-fold cross validation method, were data set first performed on training data set before any attribute reduction methods. It is shown that Jrip has the highest accuracy rate with 86.96% followed by MLP with 85.7% and C4.5 with 84.67%. As well, when test data set supplied to the model a better accuracy is achieved as the following: Jrip has the highest accuracy rate with 92.6%, followed by 89.4% for both MLP and 85.53% for C4.5.

The comparison of the data mining algorithms performed after applying PCA on Stroke data, showed that C4.5 has the highest accuracy on the test data set (95.25%).

TABLE I. Classifiers Performance using Accuracy

| Algorithms | Training set (10 fold-cross validation) | Test set |
|---|---|---|
| Performance on raw data | | |
| MLP | 85.70% | 89.40% |
| Jrip | 86.96% | 92.60% |
| C4.5 | 84.77% | 85.50% |
| Performance on after principle component analysis | | |
| MLP | 89.85% | 94.42% |
| Jrip | 88.81% | 93.18% |
| C4.5 | 88.81% | 95.25% |

Generally, it can be seen that C4.5 and Jrip are the highest classifiers in name of accuracy after PCA on the unseen data set (test data set), Table I.

## V. Discussion

In an attempt to use stroke data for the predication of stroke patients, the difference between the process on raw data, and after principle component analysis were examined. The result obtained in this research confirmed the benefit of PCA.

The technique can be used in collaboration with C4.5, Jrip, and MLP as a new framework in identifying new stroke patient. This framework works by reducing number of attributes (variables) to the optimal number using Csf subset evaluation, followed by PCA and then supplies the new data set to the three chosen algorithms. The research found that the accuracy of this approach is approximately 95% (for C4.5 algorithm), compared to un-processed data. The proposed approach showed an improvement of classification accuracy on the test data by 9.72%, 0.58%, 5.02% for J48, Jrip, and MLP respectively. Finally, based on this experiment the highest achieved accuracy was for C4.5 by 95.25%.

The results of this study showed that among the important lab test abnormalities to diagnose stroke, creatine kinase-MB (CKMB) came first, followed by lymph auto, eGFR, and HbA1C. CKMB is a test used to determine if the elevation of creatine kinase is due to heart muscle damage or skeletal muscle damage [23]. Lymph auto, is a lab test that measure white blood cells to exclude any immune system diseases [23]. The estimated glomerular filtration rate (eGFR), is a lab test used to screen renal function and evaluate it [23]. Finally, Hemoglobin A1c (HbA1C) used mainly to diagnose diabetes militias by measuring the average blood glucose level through three months periods [23]. In addition, the result of attribute ranking using Information gain attribute evolution algorithm showed that patient receiving following medication has high risk to develop stroke: Atorvastatin (lipid-lowering medication) [3], Amlodipine (hypertension medication) [3], Levetiracetam (anticonvulsant medication) [3], Metformin (diabetes militias medication) [3], Aspirin (antiplatelet medication) [3], Clopidogrel (antiplatelet medication) [3].

This means that patient who develop heart diseases (hypertension mainly), immunity diseases, diabetes militias, kidney diseases, hyperlipidemia, epilepsy, or blood (platelets) disorders, has a higher probability to develop stroke.

## VI. Conclusion

Health care organizations have gained big benefit from data mining in name of big data analysis and decision support system. In this research, stroke patient data has been collected from kamc-ngha that ended up with 17 attributes rather than class attribute.

## Acknowledgment

This work was in part supported by MITACS and, NSERC granting agencies as well as the cfi new opportunities program. The King Abdulaziz City for Science and Technology (KACST) grant supported this research paper (grant# KACST 37-1895 أ ط). In addition, we would like to thank King Abdulaziz medical city and King Abdullah International Medical Research Center for their support and for providing us with the required data.

## Disclosures

None of the authors have any competing interests.

## References

[1] Jiawei Han, et al. Data Mining Concepts and Techniques (2011). Third edition; P84-88.

[2] The American Heart and Stroke Association.

[3] Edward C Jauch, et al. Ischemic Stroke. Medscape.

[4] Guidelines for the early management of patients with acute ischemic stroke. A guideline for healthcare professionals from the American Heart Association/American Stroke Association. March 2013.

[5] Mozaffarian D, et al. on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics (2015 update) a report from the American Heart Association [published online ahead of print December 17, 2014]. Circulation. doi: 10.1161/CIR.0000000000000152

[6] World health organization. The top 10 causes of death. http://www.who.int/mediacentre/factsheets/fs310/en/ .Mayo 2014.

[7] Ziad A. Memish, et al. Burden of Disease, Injuries, and Risk Factors in the Kingdom of Saudi Arabia,1990–2010. Preventing Chronic Disease public health research, practice, and policy volume 11, e169 October 2014.

[8] Institute for Health Metrics and Evaluation. GBD Profile: Saudi Arabia. www.healthmetricsandevaluation.org. 2010.

[9] UPMC Presbyterian. A Designated Comprehensive Stroke Center what to expect: recovering from stroke. www.UPMC.com/services/stroke-institute. 2015.

[10] Lian-Yu Lin, et al. Risk factors and incidence of ischemic stroke in Taiwanese with nonvalvular atrial fibrillation—A nationwide database analysis. Atherosclerosis 217 (2011) 292–295. doi: 10.1016/j.atherosclerosis.2011.03.033. Epub 2011 Apr 5.

[11] José Rafael Romero, et al. Stroke prevention: modifying risk factors. Therapeutic Advances in Cardiovascular Disease. (2008) August; 2(4): 287–303. doi:10.1177/1753944708093847.

[12] Antonio Coca, et al. Predicting Stroke Risk in Hypertensive Patients With Coronary Artery Disease. Stroke. AHA Journals Home (2008) Feb;39(2):343-8.

[13] Tanika N. Kelly, et al. Cigarette Smoking and Risk of Stroke in the Chinese Adult Population. AHA journals. June 2008. DOI: 10.1161/STROKEAHA.107.50530.

[14] Wenbin Liang, et al. Tea Consumption and Ischemic Stroke Risk Case–Control Study in Southern China. AHA journals. July 2009. DOI: 10.1161/STROKEAHA.109.548586

[15] Fuk-hay Tang, et al. An image feature approach for computer-aided detection of ischemic stroke. Computers in Biology and Medicine 41 (2011) 529–536. doi: 10.1016/j.compbiomed.2011.05.001.

[16] A. Przelaskowskia, et al. Improved early stroke detection: Wavelet-based perception enhancement of computerized tomography exams. Computers in Biology and Medicine 37 (2007) 524 – 533. DOI: http://dx.doi.org/10.1016/j.compbiomed.2006.08.004

[17] Kartheeban Nagenthiraja, et al. Automated decision-support system for prediction of treatment responders in acute ischemic stroke. Frontiers in Neurology (2013) Volume4:Article140. doi:10.3389/fneur.2013.00140

[18] Leila Amini, et al. Prediction and Control of Stroke by Data Mining. International journal of preventive Medicine. 2013 May; 4(Suppl 2): S245–S249.

[19] Sheng-Feng Sunga, et al. Developing a stroke severity index based on administrative data was feasible using data mining techniques .Journal of Clinical Epidemiology. Volume 68, Issue 11, November 2015, Pages 1292–1300

[20] Ahmet K. Arslana, Cemil Colaka, and Ediz Sarihanb. Different Medical Data Mining Approaches Based Prediction of Ischemic Stroke. Computer Methods and Programs in Biomedicine. March 2016.

[21] D.Shanthi,,et al. Designing an Artificial Neural Network Model for the Prediction of Thromboembolic Stroke (IJBB), 2008, Volume 3. pp.10-18.

[22] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[23] Lab test online. ©2001 - 2017 by American Association for Clinical Chemistry. https://labtestsonline.org/.

[24] Tjortjis C, et al. Using T3, an improved decision tree classifier, for mining stroke-related medical data. Methods of Information in Medicine (2007) ;46(5):523-9. Doi: http://dx.doi.org/10.1160/ME0317.

[25] A. Sudha. et al. Effective analysis and predictive model of stroke disease using classification methods. international journal for computer application (0975-8887). April 2012. Voulume 43-No.14.

[26] Poonam Gupta, Rohit Miri, S.R.Tandan, Decision Tree Applied For Detecting Intrusion, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5, May – 2013 ISSN: 2278-0181.

[27] Anil Rajput , Ramesh Prasad Aharwal, et al. J48 and JRIP Rules for E-Governance Data. International Journal of Computer Science and Security (IJCSS), Volume (5) : Issue (2) : 2011

[28] Multi layer Perceptron. http://neuroph.sourceforge.net/tutorials/MultiLayerPerceptron.html